# Problem Statement

# Introduction

## 1.1 Genetic Code and Codon Degeneracy

In biological system information is stored permanently in DNA in the form of 4 nucleotides A, T G and C. This information is used to produce protein which is a sequence of 20 amino acids (Table 1). The standard version of the genetic code table presents 64 nucleotide triplets (or codons) that include 61 sense codons and three stop codons. This table defines a translational mapping from 61 sense codons to 20 amino acids (Table 2). Except Met and Trp, the other eighteen amino acids are encoded by two to six different codons. Different codons that encode the same amino acid are known as synonymous codons.

**Table 1: List of Amino acids**

|  |  |
| --- | --- |
| **Amino Acid**  **One Letter Code** | **Amino Acid**  **Full name/ meaming** |
| **A** | **Alanine** |
| **B** | **Aspartic acid or Asparagine** |
| **C** | **Cysteine** |
| **D** | **Aspartic acid** |
| **E** | **Glutamic acid** |
| **F** | **Phenylalanine** |
| **G** | **Glycine** |
| **H** | **Histidine** |
| **I** | **Isoleucine** |
| **K** | **Lysine** |
| **L** | **Leucine** |
| **M** | **Methionine** |
| **N** | **Asparagine** |
| **O** | **Pyrrolysine** |
| **P** | **Proline** |
| **Q** | **Glutamine** |
| **R** | **Arginine** |
| **S** | **Serine** |
| **T** | **Threonine** |
| **U** | **Selenocysteine** |
| **V** | **Valine** |
| **W** | **Tryptophan** |
| **Y** | **Tyrosine** |
| **Z** | **Glutamic acid or Glutamine** |
| **X** | **any** |
| **\*** | **translation stop** |

Table 1: Genetic code table

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Second Letter | | | | | | | | | | | |
| First Letter |  | U | | C | | A | | G | |  | Third Letter |
| U | UUU | Phe | UCU | Ser | UAU | Tyr | UGU | Cys | U |
| UUC | UCC | UAC | UGC | C |
| UUA | Leu | UCA | UAA | Stop | UGA | Stop | A |
| UUG | UCG | UAG | UGG | Trp | G |
| C | CUU | Leu | CCU | Pro | CAU | His | CGU | Arg | U |
| CUC | CCC | CAC | CGC | C |
| CUA | CCA | CAA | Gln | CGA | A |
| CUG | CCG | CAG | CGG | G |
| A | AUU | Ile | ACU | Thr | AAU | Asn | AGU | Ser | U |
| AUC | ACC | AAC | AGC | C |
| AUA | ACA | AAA | Lys | AGA | Arg | A |
| AUG | Met | ACG | AAG | AGG | G |
| G | GUU | Val | GCU | Ala | GAU | Asp | GGU | Gly | U |
| GUC | GCC | GAC | GGC | C |
| GUA | GCA | GAA | Glu | GGA | A |
| GUG | GCG | GAG | GGG | G |

1.2 Convert Codon sequence to AA sequence

>Ecol\_K12\_MG1655:|2706|ompC

atgaaagttaaagtactgtccctcctggtcccagctctgctggtagcaggcgcagcaaacgctgctgaagtttacaacaaagacggcaacaaattagatctgtacggtaaagtagacggcctgcactatttctctgacaacaaagatgtagatggcgaccagacctacatgcgtcttggcttcaaaggtgaaactcaggttactgaccagctgaccggttacggccagtgggaatatcagatccagggcaacagcgctgaaaacgaaaacaactcctggacccgtgtggcattcgcaggtctgaaattccaggatgtgggttctttcgactacggtcgtaactacggcgttgtttatgacgtaacttcctggaccgacgtactgccagaattcggtggtgacacctacggttctgacaacttcatgcagcagcgtggtaacggcttcgcgacctaccgtaacactgacttcttcggtctggttgacggcctgaactttgctgttcagtaccagggtaaaaacggcaacccatctggtgaaggctttactagtggcgtaactaacaacggtcgtgacgcactgcgtcaaaacggcgacggcgtcggcggttctatcacttatgattacgaaggtttcggtatcggtggtgcgatctccagctccaaacgtactgatgctcagaacaccgctgcttacatcggtaacggcgaccgtgctgaaacctacactggtggtctgaaatacgacgctaacaacatctacctggctgctcagtacacccagacctacaacgcaactcgcgtaggttccctgggttgggcgaacaaagcacagaacttcgaagctgttgctcagtaccagttcgacttcggtctgcgtccgtccctggcttacctgcagtctaaaggtaaaaacctgggtcgtggctacgacgacgaagatatcctgaaatatgttgatgttggtgctacctactacttcaacaaaaacatgtccacctacgttgactacaaaatcaacctgctggacgacaaccagttcactcgtgacgctggcatcaacactgataacatcgtagctctgggtctggtttaccagttctaa

>Ecol\_K12\_MG1655:|2706|ompC

MKVKVLSLLVPALLVAGAANAAEVYNKDGNKLDLYGKVDGLHYFSDNKDVDGDQTYMRLGFKGETQVTDQLTGYGQWEYQIQGNSAENENNSWTRVAFAGLKFQDVGSFDYGRNYGVVYDVTSWTDVLPEFGGDTYGSDNFMQQRGNGFATYRNTDFFGLVDGLNFAVQYQGKNGNPSGEGFTSGVTNNGRDALRQNGDGVGGSITYDYEGFGIGGAISSSKRTDAQNTAAYIGNGDRAETYTGGLKYDANNIYLAAQYTQTYNATRVGSLGWANKAQNFEAVAQYQFDFGLRPSLAYLQSKGKNLGRGYDDEDILKYVDVGATYYFNKNMSTYVDYKINLLDDNQFTRDAGINTDNIVALGLVYQFStop

1.2 Codon Usage Bias

Even though synonymous codons encode the same amino acids, it has been found that these codons are not used uniformly in organisms. Some synonymous codons are used more frequently than the other synonymous codons. This phenomenon is known as codon usage bias (CUB), which is variable across different organisms as well as among genes within a genome. For example in ompC gene of E. coli codon usage is given as in table 3.

**Table 3: Codon Usage in ompC Gene**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **AA** | **Codon** | **Codon Count** | **AA** | **Codon** | **Codon Count** | **AA** | **Codon** | **Codon Count** | **AA** | **Codon** | **Codon Count** |
| Phe | UUU | 2 | Ser | UCU | 6 | Tyr | UAU | 5 | Cys | UGU | 0 |
|  | UUC | 17 |  | UCC | 8 |  | UAC | 24 |  | UGC | 0 |
| Leu | UUA | 1 |  | UCA | 0 | TER | UAA | 1 | TER | UGA | 0 |
|  | UUG | 0 |  | UCG | 0 |  | UAG | 0 | Trp | UGG | 4 |
|  | CUU | 1 | Pro | CCU | 0 | His | CAU | 0 | Arg | CGU | 12 |
|  | CUC | 1 |  | CCC | 0 |  | CAC | 1 |  | CGC | 1 |
|  | CUA | 0 |  | CCA | 3 | Gln | CAA | 1 |  | CGA | 0 |
|  | CUG | 24 |  | CCG | 1 |  | CAG | 20 |  | CGG | 0 |
| Ile | AUU | 0 | Thr | ACU | 12 | Asn | AAU | 0 | Ser | AGU | 1 |
|  | AUC | 10 |  | ACC | 12 |  | AAC | 32 |  | AGC | 2 |
|  | AUA | 0 |  | ACA | 0 | Lys | AAA | 17 | Arg | AGA | 0 |
| Met | AUG | 4 |  | ACG | 0 |  | AAG | 0 |  | AGG | 0 |
| Val | GUU | 12 | Ala | GCU | 18 | Asp | GAU | 9 | Gly | GGU | 29 |
|  | GUC | 2 |  | GCC | 0 |  | GAC | 23 |  | GGC | 19 |
|  | GUA | 9 |  | GCA | 8 | Glu | GAA | 11 |  | GGA | 0 |
|  | GUG | 2 |  | GCG | 3 |  | GAG | 0 |  | GGG | 0 |

## 1.3 Measures of Codon Usage Bias

In the past thirty years several methodologies have been proposed to measure codon usage bias in organisms. Here is a measure of CUB known as Effective number of codons (Nc).

### 1.3.1 Effective number of codons (Nc)

Effective number of codons (Nc, Wright 1990) is a method of measuring overall degree of codon usage bias in a gene. It measures biasness from equal usage of alternative synonymous codons in a gene. It takes the value of 61(for standard genetic code) when all the synonymous codons are used with equal frequency, and its value decreases to as low as 20 when synonymous codon usages become highly biased.

For an amino acid *AA* with degeneracy *k*, i.e. with *k* number of synonymous codons, each with counts *n1*, *n2*,…, *nk*, *n* as the total codon count, and *pi*= *ni* / *n*, effective number of codons of the *AA* is calculated as follows:



* Equation 1



Where - Equation 2

For Example:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Hypothetical AA | No of codons | | | | | | *n* | *p* | | | | | |  | *FAA* | *NcAA* |
| X | 0 | 0 | 0 | 0 | 0 | 125 | 125 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |  | 1.00 | 1.00 |
| Y | 150 | 150 | 150 | 150 | 150 | 150 | 900 | 0.17 | 0.17 | 0.17 | 0.17 | 0.17 | 0.17 |  | 0.17 | 6.03 |

Finally effective number of codons of the gene is computed as



* Equation 3

**An Example to calculate Nc value of a gene ompC**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **AA** | **No of codons** | | | | | | **n** | **pi** | | | | | | **pi2** | | | | | | **ΣPi2** | **Fk** | **1/Fk** | **Nc** |
| Met | 4 |  |  |  |  |  | 4 | 1.00 |  |  |  |  |  | 1.00 |  |  |  |  |  | 1.00 | 1.00 | 1.00 |  |
| Trp | 4 |  |  |  |  |  | 4 | 1.00 |  |  |  |  |  | 1.00 |  |  |  |  |  | 1.00 | 1.00 | 1.00 |  |
| Phe | 2 | 17 |  |  |  |  | 19 | 0.11 | 0.89 |  |  |  |  | 0.01 | 0.80 |  |  |  |  | 0.81 | 0.81 | 1.23 |  |
| Tyr | 5 | 24 |  |  |  |  | 29 | 0.17 | 0.83 |  |  |  |  | 0.03 | 0.68 |  |  |  |  | 0.71 | 0.71 | 1.40 |  |
| His | 0 | 1 |  |  |  |  | 1 | 0.00 | 1.00 |  |  |  |  | 0.00 | 1.00 |  |  |  |  | 1.00 | 1.00 | 1.00 |  |
| Gln | 1 | 20 |  |  |  |  | 21 | 0.05 | 0.95 |  |  |  |  | 0.00 | 0.91 |  |  |  |  | 0.91 | 0.91 | 1.10 |  |
| Asn | 0 | 32 |  |  |  |  | 32 | 0.00 | 1.00 |  |  |  |  | 0.00 | 1.00 |  |  |  |  | 1.00 | 1.00 | 1.00 |  |
| Lys | 17 | 0 |  |  |  |  | 17 | 1.00 | 0.00 |  |  |  |  | 1.00 | 0.00 |  |  |  |  | 1.00 | 1.00 | 1.00 |  |
| Asp | 9 | 23 |  |  |  |  | 32 | 0.28 | 0.72 |  |  |  |  | 0.08 | 0.52 |  |  |  |  | 0.60 | 0.60 | 1.68 |  |
| Glu | 11 | 0 |  |  |  |  | 11 | 1.00 | 0.00 |  |  |  |  | 1.00 | 0.00 |  |  |  |  | 1.00 | 1.00 | 1.00 |  |
| Cys | **0** | **0** |  |  |  |  | 0 | **###** | **###** |  |  |  |  | **####** | **###** |  |  |  |  | **NA** | **NA** | **NA** |  |
| Ile | 0 | 10 | 0 |  |  |  | 10 | 0.00 | 1.00 | 0.00 |  |  |  | 0.00 | 1.00 | 0.00 |  |  |  | 1.00 | 1.00 | 1.00 |  |
| Val | 12 | 2 | 9 | 2 |  |  | 25 | 0.48 | 0.08 | 0.36 | 0.08 |  |  | 0.23 | 0.01 | 0.13 | 0.01 |  |  | 0.37 | 0.37 | 2.68 |  |
| Pro | 0 | 0 | 3 | 1 |  |  | 4 | 0.00 | 0.00 | 0.75 | 0.25 |  |  | 0.00 | 0.00 | 0.56 | 0.06 |  |  | 0.63 | 0.63 | 1.60 |  |
| Thr | 12 | 12 | 0 | 0 |  |  | 24 | 0.50 | 0.50 | 0.00 | 0.00 |  |  | 0.25 | 0.25 | 0.00 | 0.00 |  |  | 0.50 | 0.50 | 2.00 |  |
| Ala | 18 | 0 | 8 | 3 |  |  | 29 | 0.62 | 0.00 | 0.28 | 0.10 |  |  | 0.39 | 0.00 | 0.08 | 0.01 |  |  | 0.47 | 0.47 | 2.12 |  |
| Gly | 29 | 19 | 0 | 0 |  |  | 48 | 0.60 | 0.40 | 0.00 | 0.00 |  |  | 0.37 | 0.16 | 0.00 | 0.00 |  |  | 0.52 | 0.52 | 1.92 |  |
| Leu | 1 | 0 | 1 | 1 | 0 | 24 | 27 | 0.04 | 0.00 | 0.04 | 0.04 | 0.00 | 0.89 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.79 | 0.79 | 0.79 | 1.26 |  |
| Ser | 6 | 8 | 0 | 0 | 1 | 2 | 17 | 0.35 | 0.47 | 0.00 | 0.00 | 0.06 | 0.12 | 0.12 | 0.22 | 0.00 | 0.00 | 0.00 | 0.01 | 0.36 | 0.36 | 2.75 |  |
| Arg | 12 | 1 | 0 | 0 | 0 | 0 | 13 | 0.92 | 0.08 | 0.00 | 0.00 | 0.00 | 0.00 | 0.85 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.86 | 0.86 | 1.17 | **27.90** |